Лекция 8. Аппаратное обеспечение для BigData

Если спросить специалиста, какая конфигурация оборудования лучше всего подходит для выполнения добычи данных, единственный подходящий ответ на этот вопрос заключается в том, что это зависит от того, что вы пытаетесь сделать. Существует целый ряд соображений, которые необходимо взвесить при принятии решения о том, как создать подходящую вычислительной среды для анализа больших данных.

Хранилище (диск)

Хранение данных — это первое, что приходит на ум при упоминании темы больших данных. Именно хранение данных позволяет нам вести учет истории, чтобы использовать ее для определения того, что может произойти в будущем.

Традиционный жесткий диск состоит из пластин, которые представляют собой диски, покрытые намагниченной пленкой, позволяющей кодировать 1 и 0, составляющие данные. Шпиндели, которые вращают вертикально расположенные пластины, являются критически важной частью рейтинговых жестких дисков, поскольку шпиндели определяют скорость вращения пластин и, следовательно, скорость чтения и записи данных. Каждая пластина имеет одну головку привода; они движутся в унисон, так что только одна головка считывает данные с определенной пластины.

Эта механическая операция является очень точной и одновременно очень медленной по сравнению с другими компонентами компьютера. Это может значительно увеличить время, необходимое для решения высокопроизводительных задач по поиску данных.

Для борьбы с недостатком дисковых скоростей широкое распространение получили дисковые массивы[1], которые обеспечивают более высокую пропускную способность. Максимальная пропускная способность дискового массива для одной системы от внешних подсистем хранения данных находится в диапазоне от 1 до 6 гигабайт (ГБ) в секунду (ускорение скорости доступа к данным в 10-50 раз).

Еще одно изменение в дисковых накопителях в ответ на эру больших данных заключается в том, что их емкость за последние 10 лет увеличивалась на 50%-100% в год. Кроме того, цены на дисковые массивы оставались практически неизменными, что означает, что цена за терабайт (ТБ) снижалась на половину в год. Увеличение емкости дисковых накопителей не сопровождалось увеличением возможности передачи данных на/ с диска, которая увеличивалась всего на 15-20 % в год. Для примера, в 2008 году объем типичного серверного диска составлял 500 ГБ, а скорость передачи данных - 98 мегабайт в секунду (МБ/с). Весь диск можно было передать примерно за 85 минут (500 ГБ = 500 000 МБ/98 МБ/сек). В 2013 году появились диски емкостью 4 ТБ со скоростью передачи данных 150 МБ/с, но для передачи всего диска потребуется около 440 минут. Если учесть, что объем данных удваивается каждые несколько лет, проблема очевидна. Необходимы более быстрые диски.

---

[1] Дисковый массив - это специализированное аппаратное хранилище, которое обеспечивает большую емкость хранения и доступ к данным благодаря своей специализированной реализации. NetApp и EMC - два основных производителя дисковых массивов. Преимущества могут заключаться в параллельном доступе к дискам, который возникает при одновременном чтении данных с нескольких дисков для решения одной и той же задачи. Это является преимуществом только в том случае, если программное обеспечение может использовать преимущества такой конфигурации дисковых накопителей.

Твердотельные устройства (SSD) - это дисковые накопители без диска или каких-либо движущихся частей. Их можно рассматривать как стабильную память, а скорость чтения данных с них может легко превышать 450 МБ/с. Для сред добычи данных умеренного размера твердотельные накопители и их превосходная пропускная способность могут кардинально изменить время решения проблемы. Массивы SSD также доступны, но SSD по-прежнему стоят значительно дороже, чем жесткие диски (HDD), в расчете на единицу емкости. Массивы SSD ограничены той же пропускной способностью внешнего хранилища, что и массивы HDD. Поэтому, хотя твердотельные накопители могут решить проблему добычи данных за счет сокращения общего времени на чтение и запись данных, перевод всего хранилища на SSD может оказаться непомерно дорогим. В этом случае необходимы гибридные стратегии, использующие различные типы устройств. Еще один момент - размер дисков, которые приобретаются для аналитических нагрузок. Диски меньшего размера имеют более быстрое время доступа.

Исторически сложилось так, что только некоторые аналитические программы могли использовать дополнительную память для увеличения объема памяти путем записи промежуточных результатов на диск. Это расширяло размер решаемой задачи, но приводило к увеличению времени выполнения. Время выполнения увеличивалось не только из-за дополнительной загрузки данных, но и из-за более медленного доступа к данным при чтении промежуточных результатов с диска вместо чтения их из памяти. Для типичной настольной или небольшой серверной системы доступ к данным на устройства хранения, особенно запись на устройства хранения, является мучительно медленным. Один поток выполнения аналитического процесса может легко потреблять 100 МБ/с, а доминирующим типом доступа к данным является последовательное чтение или запись. Типичная рабочая станция высокого класса оснащена диском SAS со скоростью вращения 15 000 об/мин; диск вращается со скоростью 15 000 оборотов в минуту и использует технологию SAS для чтения и записи данных со скоростью от 100 до 150 МБ/сек. Это означает, что одно или два ядра могут потреблять всю доступную пропускную способность диска. Это также означает, что в современной системе с большим количеством ядер большой процент ресурсов центрального процессора (ЦП) будет простаивать при выполнении многих операций по поиску данных; это не отсутствие необходимых вычислительных ресурсов, а несоответствие между дисками, памятью и ЦП.

Центральный процессор

Термин "центральный процессор" имеет два значения в компьютерном оборудовании. CPU используется для обозначения пластикового и стального корпуса, в котором находятся все основные элементы компьютера. Сюда входят блок питания, материнская плата, периферийные карты и так далее. Другое значение CPU - это вычислительная микросхема, расположенная внутри пластикового и стального корпуса. В данной лекции под процессором подразумевается микросхема.

В 1980-х и 1990-х годах скорость центрального процессора резко возросла. Скорость процессора увеличивалась настолько, что однопоточные программные приложения выполнялись почти в два раза быстрее на новых версиях процессоров по мере их появления. Ускорение процессора было описано Гордоном Муром, соучредителем компании Intel, в знаменитом законе Мура, который заключается в том, что каждые два года количество транзисторов и интегральных схем, которые можно разместить на определенной площади, удваивается, и поэтому инструкции могут выполняться с удвоенной скоростью. Эта тенденция удвоения скорости процессоров продолжилась в 1990-х годах, когда инженеры Intel заметили, что если тенденция

удвоения продолжится, то к 2010 году тепло, выделяемое этими чипами, будет таким же горячим, как солнце. В начале 2000-х годов "бесплатный обед" по закону Мура закончился, по крайней мере, с точки зрения скорости обработки данных. Скорость (частота) процессоров остановилась, и компьютерные компании искали новые способы увеличения производительности. Векторные блоки, присутствующие в ограниченной форме в x86 начиная с инструкций Pentium MMX, становились все более важными для достижения производительности и приобретали дополнительные возможности, такие как плавающая точка с одинарной, а затем и двойной точностью. В начале 2000-х годов производители чипов также обратились к добавлению дополнительных потоков выполнения в свои чипы. Эти многоядерные чипы представляли собой уменьшенные версии многопроцессорных суперкомпьютеров, в которых ядра совместно использовали такие ресурсы, как кэш-память. Количество ядер, расположенных на одном чипе, со временем увеличилось; сегодня многие серверные машины предлагают два шестиядерных процессора.

По сравнению с доступом к данным на жестком диске, доступ процессора к памяти быстрее, чем разгоняющаяся пуля; типичный доступ находится в диапазоне от 10 до 30 ГБ/сек. Все остальные компоненты компьютера мчатся наперегонки, чтобы не отстать от ЦП.

Графический процессор

Графический процессор (GPU) получил широкую известность как неиспользуемый вычислительный ресурс, который может сократить время выполнения задач по поиску данных и других аналитических задач за счет распараллеливания вычислений. GPU уже есть в каждом настольном компьютере в мире. В начале 2000-х годов графические процессоры включились в вычислительную игру. Обработка графики значительно эволюционировала по сравнению с ранними текстовыми дисплеями первых настольных компьютеров. Стремление к улучшению графики было вызвано потребностями промышленности в инструментах визуализации. Одним из примеров является использование инженерами трехмерного (3D) программного обеспечения автоматизированного проектирования (CAD) для создания прототипов новых конструкций еще до их создания. Еще более важным фактором развития вычислений на GPU стала индустрия потребительских видеоигр, в которой наблюдаются тенденции изменения цен и производительности, аналогичные остальным отраслям потребительской вычислительной техники. Неустанное стремление к повышению производительности при меньших затратах позволило среднему пользователю получить неслыханную производительность как на CPU, так и на GPU.

Для создания анимации обработка трехмерной графики должна обрабатывать миллионы или миллиарды трехмерных треугольников в трехмерных сценах несколько раз в секунду. Размещение и раскраска всех этих треугольников в их трехмерном окружении требует огромного количества одинаковых вычислений. Первоначально 3D-графика выполнялась с помощью фиксированного конвейера рендеринга, который принимал информацию о 3D-сцене и превращал ее в пиксели, которые можно было представить пользователю в видео или на экране. Этот фиксированный конвейер был реализован аппаратно, причем различные части графического процессора выполняли различные части задачи по превращению треугольников в пиксели. В начале 2000-х годов этот фиксированный конвейер был постепенно заменен обобщенными программными шейдерами, которые представляли собой мини-программы, выполняющие операции предыдущего фиксированного аппаратного конвейера.

С помощью этих шейдеров специалисты по высокопроизводительным вычислениям заметили, что координаты и цвета с плавающей запятой могут выглядеть

очень похоже на координаты в задачах по физике или химии, если посмотреть на них правильно. Более хардкорные хакеры начали создавать графические задачи, которые выглядели как бессмыслица, за исключением того, что вычисления, лежащие в их основе, очень быстро решали трудные задачи. Был замечен рост производительности, и были разработаны вычислительные платформы, которые использовали GPU для выполнения неграфических вычислений. Эти вычисления относятся к тому же типу, который необходим для data mining.

Графические процессоры – это "зеленое поле". Исторически возможность разработки кода для работы на GPU была ограниченной и дорогостоящей. За последние несколько лет интерфейсы программирования для разработки программного обеспечения, использующего преимущества GPU, значительно улучшились. Программное обеспечение только начало использовать преимущества GPU, и пройдет еще несколько лет, прежде чем вычисления, необходимые для добычи данных, будут эффективно передаваться на GPU для выполнения. Когда это время наступит, ускорение многих типов задач по добыче данных сократится с часов до минут и с минут до секунд.

Память

Память, или память с произвольным доступом (RAM), как ее обычно называют, является важнейшим и часто недооцениваемым компонентом при создании платформы для добычи данных. Память является посредником между хранением данных и обработкой математических операций, выполняемых центральным процессором. Память является волатильной, что означает, что если она теряет питание, то данные, хранящиеся в ней, теряются.

В 1980-х и 1990-х годах разработка алгоритмов добычи данных была сильно ограничена как памятью, так и процессором. Ограничение по памяти было связано с 32-разрядными операционными системами, которые позволяли использовать только 4 ГБ памяти. Это ограничение фактически означало, что ни одна задача по поиску данных, требующая более 4 ГБ памяти[2] (за вычетом программного обеспечения и операционной системы, работающей на машине), не могла быть решена с использованием только памяти. Это очень важно, поскольку пропускная способность памяти обычно составляет от 12 до 30 ГБ/сек, а самого быстрого хранилища - всего около 6 ГБ/сек, причем пропускная способность большинства хранилищ гораздо меньше.

Примерно в 2004 году аппаратное обеспечение (Intel и AMD) поддерживало 64-битные вычисления. В то же время, когда операционные системы стали поддерживать большие объемы памяти, фактическая цена памяти резко снизилась. В 2000 году средняя цена 1 МБ оперативной памяти составляла $1,12. В 2005 году средняя цена составляла $0,185, а в 2010 году - $0,0122.

Благодаря поддержке 64-разрядных вычислительных систем, которые могут использовать до 8 ТБ памяти, и снижению цен на память, стало возможным создавать платформы для добычи данных, которые могли хранить всю задачу добычи данных в памяти. Это, в свою очередь, позволяет получать результаты за конкретное время.

Алгоритмы добычи данных часто требуют, чтобы все данные и вычисления производились в памяти. Без внешнего хранения данных увеличение виртуального и реального адресного пространства, а также резкое снижение цен на память создали возможность решать многие проблемы data mining, которые ранее были

[2] Наибольшее целочисленное значение, которое 32-разрядные операционные системы могут использовать для адресации или ссылки на память, составляет 2 32 -1, или 3,73 ГБ памяти

неосуществимы. Чтобы проиллюстрировать этот пример, рассмотрим задачу прогнозного моделирования, в которой используется алгоритм нейронной сети. Нейронная сеть будет выполнять итерационную оптимизацию для поиска наилучшей модели. Для каждой итерации она должна будет считывать данные один раз. Нейронные сети нередко совершают тысячи проходов через данные, чтобы найти оптимальное решение.

Если эти проходы выполняются в памяти со скоростью 20 ГБ/сек против 1 ГБ/сек на диске, то проблема, на решение которой в памяти уходит всего 10 секунд, на диске будет решаться более 3 минут. Если этот сценарий повторяется часто, производительность специалиста по добыче данных резко падает. В дополнение к производительности человеческого капитала, если процессы добычи данных зависят от дискового хранилища, вычисления будут выполняться во много раз дольше. Чем дольше длится процесс, тем выше вероятность того, что произойдет какой-то аппаратный сбой. Такие сбои, как правило, не поддаются восстановлению, и весь процесс приходится запускать заново.

Скорость памяти увеличивалась гораздо более умеренными темпами, чем скорость процессоров. Скорость памяти увеличилась в 10 раз по сравнению со скоростью процессора, которая увеличилась в 10 000 раз. Пропускная способность дисковых накопителей растет еще медленнее, чем память. В результате алгоритмы добычи данных преимущественно хранят все структуры данных в памяти и перешли к распределенным вычислениям, чтобы увеличить объем вычислений и памяти. Пропускная способность памяти обычно находится в диапазоне от 12 до 30 ГБ/с, и память очень дешева. Пропускная способность систем хранения данных достигает 6 ГБ/с и является очень дорогой. Гораздо дешевле развернуть набор товарных систем со здоровым объемом памяти, чем покупать дорогие высокоскоростные дисковые системы хранения данных. Современные серверные системы обычно оснащаются от 64 до 256 ГБ памяти. Для получения быстрых результатов необходимо учитывать размер памяти.

Сеть

Сеть – это единственный аппаратный компонент, который всегда является внешним по отношению к компьютеру[3]. Это механизм связи компьютеров с другими компьютерами.

Скорость сети должна быть фактором только для распределенной вычислительной среды. В случае отдельного компьютера (рабочей станции или сервера) данные, память и центральный процессор должны быть локальными, и производительность аналитической задачи не будет зависеть от скорости сети. Стандартным сетевым соединением для кластера аналитических вычислений является 10-гигабитный Ethernet (10 GbE), который имеет верхнюю границу скорости передачи данных в 4 гигабайта в секунду (ГБ/сек). Эта скорость передачи данных намного медленнее, чем все остальные основные элементы, о которых шла речь. Собственные протоколы, такие как InfinibandR, обеспечивают лучшую пропускную способность, но все равно не соответствуют скорости других компонентов. По этой причине очень важно минимизировать использование сети для перемещения данных или даже для несущественной связи между различными узлами вычислительного устройства.

Именно это узкое место в скорости сети делает параллелизацию ряда алгоритмов добычи данных столь сложной задачей. Требуется значительное мастерство в программной инфраструктуре, выборе алгоритмов и окончательной реализации, чтобы

[3] Иногда хранилище может быть внешним в сети хранения данных (SAN).

эффективно и точно обработать модель с помощью одного из многих алгоритмов, не перемещая при этом данные и ограничивая связь между компьютерами. Скорость сети вашей высокопроизводительной платформы для добычи данных будет иметь большое значение, если у вас распределенная вычислительная среда. Поскольку скорость передачи данных по сети намного ниже, чем у других компонентов, вы должны учитывать сетевой компонент при оценке программных решений для добычи данных.

Аппаратные решения

Существует ряд аппаратно-программных комплексов, предоставляющих предконфигурированные решения для обработки больших данных: Aster MapReduce appliance (корпорации Teradata), Oracle Big Data appliance, Greenplum appliance (корпорации EMC, на основе решений поглощённой компании Greenplum). Эти комплексы поставляются как готовые к установке в центры обработки данных телекоммуникационные шкафы, содержащие кластер серверов и управляющее программное обеспечение для массово-параллельной обработки.

Аппаратные решения для резидентных вычислений, прежде всего, для баз данных в оперативной памяти и аналитики в оперативной памяти, в частности, предлагаемой аппаратно-программными комплексами Hana (предконфигурированное аппаратно-программное решение компании SAP) и Exalytics (комплекс компании Oracle на основе реляционной системы Timesten (англ.) и многомерной Essbase), также иногда относят к решениям из области больших данных, несмотря на то, что такая обработка изначально не является массово-параллельной, а объёмы оперативной памяти одного узла ограничиваются несколькими терабайтами.

Кроме того, иногда к решениям для больших данных относят и аппаратно-программные комплексы на основе традиционных реляционных систем управления базами данных — Netezza, Teradata, Exadata, как способные эффективно обрабатывать терабайты и эксабайты структурированной информации, решая задачи быстрой поисковой и аналитической обработки огромных объёмов структурированных данных. Отмечается, что первыми массово-параллельными аппаратно-программными решениями для обработки сверхбольших объёмов данных были машины компаний Britton Lee, впервые выпущенные в 1983 году, и Teradata (начали выпускаться в 1984 году, притом в 1990 году Teradata поглотила Britton Lee).

Аппаратные решения DAS — систем хранения данных, напрямую присоединённых к узлам — в условиях независимости узлов обработки в SN-архитектуре также иногда относят к технологиям больших данных. Именно с появлением концепции больших данных связывают всплеск интереса к DAS-решениям в начале 2010-х годов, после вытеснения их в 2000-е годы сетевыми решениями классов NAS и SAN.

Список использованных источников:

1. Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners (Wiley and SAS Business Series) 1st Edition.

2. Большие данные. URL: https://ru.wikipedia.org/wiki/%D0%91%D0%BE%D0%BB%D1%8C%D1%88%D0%B8%D0%B5_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5 (Дата обращения: 21.09.2020).